LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-JRNL-647362

Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array

A. Devault, C. Jaing, S. Gardner

December 6, 2013

Scientific Reports

Disclaimer

Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array

Alison M. Devault
Corresponding Author
McMaster Ancient DNA Centre, Department of Anthropology, McMaster University

Crystal Jaing
Email: jaing2@llnl.gov

Shea Gardner
Email: gardner26@llnl.gov

LLNL-JRNL-647362

Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array

Alison M. Devault[1], Crystal Jaing[2], Shea Gardner[2], Teresita M. Porter[3], Jacob Enk[1,3], James Thissen[2], Jonathan Allen[2], Monica Borucki[2], Sharon N. DeWitte[4], Anna N. Dhody[5], Kevin McLoughin[2], and Hendrik N. Poinar[1,3,6 *]

1. McMaster Ancient DNA Centre, Department of Anthropology, McMaster University, 1280 Main St W, Hamilton, Ontario L8S4L9, Canada
2. Lawrence Livermore National Laboratory, Livermore, CA 94551, USA
3. Department of Biology, McMaster University, 1280 Main St W, Hamilton, Ontario L8S4L8, Canada
4. Departments of Anthropology and Biological Sciences, University of South Carolina, Columbia, SC, USA
5. The College of Physicians of Philadelphia, Mütter Museum, 19 S 22nd St, Philadelphia, PA 19103, USA
6. Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, 1280 Main St W, Hamilton, Ontario L8S4L8, Canada

Abstract

Ancient human remains of paleopathological interest typically contain highly degraded DNA in which pathogenic taxa are often minority components, making sequence-based metagenomic characterization costly. Microarrays may hold a potential solution to these challenges, offering a rapid, affordable, and highly informative snapshot of microbial diversity in complex samples without the lengthy analysis and/or high cost associated with high-throughput sequencing. Their versatility is well established for modern clinical specimens, but they have yet to be applied to ancient remains. Here we report bacterial profiles of archaeological and historical human remains using the Lawrence Livermore Microbial Detection Array (LLMDA). The array successfully identified previously-verified bacterial human pathogens, including *Vibrio cholerae* (cholera) in a 19th century intestinal specimen and *Yersinia pestis* ("Black Death") in a medieval tooth, which represented only minute fractions (0.03% and 0.08% alignable sequence reads) of their respective DNA content. This demonstrates that the LLMDA can identify primary and or co-infecting bacterial pathogens obtained with high-throughput shotgun sequences, thereby serving as a rapid and inexpensive paleopathological screening tool to study health across both space and time.

Introduction

Research into the origins of infectious diseases and population health through time faces many challenges, such as biased archival records and ambiguous paleopathological skeletal indicators of actual pathogen infection levels.[1] Despite its inherent fragility, ancient DNA (aDNA) remains a highly informative paleopathological study target, having been recovered and characterized from a variety of contexts, age depths and specimen types.[2] Recently, high-throughput sequencing (HTS), often coupled with targeted enrichment (TE), has allowed for the recovery of large genomic targets from archaeological specimens, including full pathogen genomes.[3-6] However, TE-HTS is only useful when the primary pathogen(s) are known or suspected to be present, and necessarily ignores non-targeted taxa and genomic loci. This is problematic because the primary pathogenic agent in an ancient paleopathological specimen can be elusive, and furthermore the entire microbiome likely played a significant role in past human health, as it does today.[7] Therefore establishing detailed levels of commensal and co-infecting pathogens is essential for accurately reconstructing past epidemics, population health, and disease susceptibility. As such, for paleopathologists wishing to examine changes in microbial co-infection levels across space and time, more comprehensive metagenomic characterization is necessary. One way to achieve this is by sequencing amplicons of conserved loci (such as 16S rRNA) that can to a degree measure the metagenomic content of a sample. However, by design, amplicon datasets ignore potential taxonomically-informative diversity in more variable genomic regions, and for that matter can be biased by polymerase or disparate target abundances.[8,9] Metagenomic "shotgun" HTS on the other hand is arguably the most comprehensive and least biased method currently available for total microbial characterization for modern and aDNA specimens[10,11], but very deep sequencing is often required to identify pathogens confidently. While certainly powerful, both of these metagenomic approaches can be labor- and time-intensive, thereby representing significant barriers for groups that would like to thoroughly profile or screen the microbial content of large or difficult paleopathological sample sets.

One potential technological solution to this issue is the microarray, which over the past two decades has been used for the large-scale study of gene expression and genic content of simple and complex samples[12]. Microarrays are glass slides densely spotted with clusters of single-stranded synthetic oligonucleotides that are allowed to hybridize with fluorophore-labeled DNA from a sample, and the resulting fluorescence signals are interpreted to determine sequence composition and/or taxonomic content. Recently, microarrays designed specifically for characterizing the microbial content of complex samples have been successfully used (e.g. [13-18] _ENREF_13), particularly in cases where traditional clinical methods are inconclusive, time-consuming, and/or expensive.[18] Microarrays can contain up to millions of unique oligonucleotides and their use and analysis involve low processing time and cost.[15] Therefore, they potentially provide a more practical alternative to metagenomic HTS for characterizing the microbial content of paleopathological specimens. However, microarray

detection techniques have not yet been applied to aDNA extracts, which due to short fragment length and base damage may present challenges to microarray analysis.

To assess the potential value of microarrays for pathogen aDNA analysis, here we compare microbial profiles of two archaeological human specimens generated with a recently-developed pathogen detection microarray to profiles generated with standard metagenomic HTS analysis. For microarray analysis, we used the Lawrence Livermore Microbial Detection Array (LLMDA) designed by the Lawrence Livermore National Laboratory,[13] one of several array platforms developed in the last decade to identify pathogens in experimental mixtures and clinical samples.[15] The LLMDA v5 12x135K array contains probes designed from all published vertebrate-infecting pathogen genomes that target regions unique to at least the family taxonomic level. Florescence data is analyzed using a likelihood maximization algorithm to identify the combination of microbial genomes that best explains the observed signals (see Supplementary Materials for full description). The specimens we analyze here are a preserved intestinal medical sample from an 1849AD cholera victim (specimen 3090.13)[3] and a tooth from a 1348AD Black Death plague victim (specimen 8291).[5] Both specimens were previously confirmed with TE-HTS to contain their relevant pathogens, though they constitute very low levels in shotgun HTS datasets (3090.13: 0.03% alignable with bowtie to *Vibrio cholerae*, the etiological agent of cholera; 8291: 0.08% alignable to *Yersinia pestis,* the etiological agent of the Black Death, [19]). Both of these pathogens' families (Vibrionaceae and Enterobacteraceae) have probes on the LLMDA and, if within the sensitivity threshold of the LLMDA, should therefore be detectable. We specifically assess (1) whether LLMDA would detect these previously determined pathogens, (2) which additional bacteria were detectable by both LLMDA and HTS, and (3) which bacteria were detected by either LLMDA or HTS alone.

Results

Here we restrict our taxonomic profile comparisons to bacterial families since the sequencing libraries were built from DNA only and thus not appropriate for a complete viral survey. Note that the v5 12x135K LLMDA probes were derived from all complete genomic sequences from vertebrate-infecting pathogens available at the time of design (December 2011). However, as the hybridization patterns were interpreted using an updated genome database (created in April 2012), probes originally designed for one family may match newly sequenced genomes from other taxa as well. In addition, probes with weak similarity to bacterial genomes from non-vertebrate infecting families may hybridize to these genomes, so that the potential taxonomic calls are not limited to those used specifically for probe design. For the metagenomic HTS data, taxonomic assignments were identified by BLAST (blastn-megablast)[20] and MEGAN4[21] analysis against the National Center for Biotechnology Information (NCBI) RefSeq genome database[22] (October 2012). Results for both methods are given in **Table 1** and **Table S1** and a schematic comparison is provided in **Figure 1.**

Taxa detected by both LLMDA & HTS

For cholera victim 3090.13, twenty-one families were detected by both LLMDA and the 118 million BLASTed HTS reads from the sample (**Figure 1**), representing 36.8% and 40.4% of the families called by each respective method. For plague victim 8291, fifty-three families were detected by both approaches, representing 89.8% and 27.9% of the families called by LLMDA and 83 million HTS reads, respectively. These overlaps included many groups with relatively high read counts in the HTS data (e.g. Aeromonadaceae and Enterobacteriaceae for 3090.13; Burkholderiaceae, Comamonadaceae, and Pseudomonadaceae for 8291). In addition, both methods detected the previously confirmed significant pathogens to the species level. For 3090.13, 10,379 (0.009% of BLAST reads) were *V. cholerae,* and LLMDA called the family *Vibrionaceae* with *V. cholerae* chromosomal sequences at a high log odds value (4,470.7). For 8291, 1,272 (0.001% of BLAST reads) were *Y. pestis,* and LLMDA called the family Enterobacteriaceae including *Y. pestis* plasmid sequences (among other species) at a high odds value (1,640.8).

The LLMDA used here only targets groups with at least one vertebrate-infecting pathogen species; however it is able to detect families not represented by distinct probes on the array. When we considered only the families on the array for with specific probes had been designed for them, we detected 19 families in the cholera victim 3090.13 by both LLMDA and HTS, representing 41.3% and 79.2% of the families called by each respective method and 46 families for the plague victim 8291, representing 92.0% and 56.8% of the families with probes on the array (LLMDA and HTS).

Taxa detected by only one method

BLAST analyses of HTS reads identified many families that were not detected by LLMDA analyses (for cholera victim 3090.13, n = 10, 32.3% of all HTS; for plague victim 8291, n = 137, 72.1% of all HTS), such as Neisseriaceae and Shewanellaceae in sample 3090.13, Cellulomonadaceae and Rhizobiaceae in sample 8291, and Fusobacteriaceae and Peptostreptococcaceae in both samples. Likewise, LLMDA analysis identified many families that HTS-MEGAN4 did not (for 3090.13, n = 36, 63.1% of all LLMDA; for 8291, n = 6, 10.2% of all LLMDA). When excluding taxonomic groups without probes designed for them on the array, that left only 5 families detected by HTS (20.8%) for sample 3090.13 and 35 (43.2%) for 8291 and 27 families for 3090.13 (58.7% of all LLMDA) and 4 for 8291 (8.0%) only detected via LLMDA.

Discussion

Figure 2 displays the MEGAN4 output of the NCBI taxonomy for all taxa identified with BLAST analysis of the HTS data and whether they were also detected with LLMDA. Overall, the LLMDA profiles reflect the major HTS-identified components well. Not only were the previously-identified pathogen

families detected via both methods, but a number of major environmental, microbiomic and pathogenic taxa were identified to at least the order level (e.g., Actinomycetales, Bacilliales, Clostridiales, or Rhizobiales). While promising, a number of disparities between the profiles generated by each method encourage further investigation into their origin (discussed below).

When comparing metagenomic profiles generated by each method, it is important to be aware of the fundamental differences in their taxonomic identification strategies. For the analysis of BLAST output from HTS data, default parameters in MEGAN4 require five sequence reads to assign a taxon as being present; furthermore, the reads do not have to be assigned to the same species for family-level calls.[21] MEGAN4 also gives equal weight to read mappings that are concentrated in narrow regions of a target genome, which are inherently less specific as indicators of the target's presence. A common possible scenario leading to false positive taxon assignments could occur in both HTS and microarray analysis, when reads or probes map to ribosomal RNA or housekeeping genes that are relatively conserved between related taxa. Microarray probes can be designed to avoid these conserved regions, but in general sequence reads mapping to such regions are not filtered out in metagenomic analysis. Therefore, BLAST/MEGAN4 analysis of HTS data emphasizes sensitivity at the expense of specificity.

The CLiMax algorithm used for LLMDA analysis requires that a family satisfy more stringent criteria to be considered present. The initial CLiMax analysis is performed at the target genome level rather than the family level; for a target to be called present, a minimum of 4 probes or 20% of the probes designed against the target (whichever is larger) must have intensities above an array-specific significance threshold. In addition, targets for which the high intensity probes are concentrated in narrow genomic regions are filtered out as potential false positives (see Supplementary Appendix for description of methods). When this filtering is removed, or if the minimum probe criteria are relaxed, CLiMax predicts the presence of several previously undetected families (data not shown). However, our previous experiments in which the LLMDA was hybridized to samples of known microbial content indicate that stringent filtering is necessary to avoid false positives.[13] Therefore, the CLiMax analysis is much more conservative in its predictions than BLAST/MEGAN4 analysis, emphasizing specificity over sensitivity.

Several taxa detected with HTS were not detected with LLMDA. Many of these are unsurprising, as no probes designed from their genomes were on the array. However for those taxa that were used for array probe design, one possibility is that the LLMDA is not as sensitive as HTS at these sequencing depths: in plague victim 8291, taxa not detected with LLMDA had significantly fewer HTS reads than those that were (two-tailed, unequal variance Student's t-test, $p = 0.004$; **Figure 3a**), though this relationship is much weaker for cholera victim 3090.13 ($p = 0.152$). Furthermore, several taxa with relatively high read counts and with probes designed on the array were surprisingly not called (e.g., Sphingomonadaceae in sample 8291; Peptostreptococcaceae in both samples). That said, in the

majority of cases where a family with probes designed on the array was declared present by BLAST/MEGAN4 analysis but not called with LLMDA, a closely-related taxon was called (e.g., in both samples, Clostridiaceae was called though its close relative Peptostreptococcaceae was not).

To better understand the data used by MEGAN4 to call family Peptostreptococcaceae as present, we examined the gene, rRNA, and other feature annotations for the mapped read positions in *Clostridium difficile* strain 630 (RefSeq accession NC_009089.1), one of the fully sequenced genomes in this family. Notably, 915 of 1328 (69%) reads mapped to this genome from cholera victim 3090.13 and 146 of 319 (46%) from plague victim 8291 were within rRNA genes. Since rRNA genes only cover 1.1% of the *C. difficile* 630 genome, these read counts are far larger than would be expected by chance alone. Consequently, we suspect that a large part of the data used by MEGAN4 to call this family as present is based on reads that map to highly conserved genes, and could also support the presence of a related taxon. Although a detailed analysis of MEGAN4 performance is beyond the scope of this study, our preliminary results suggest that its relative nonspecificity could underlie some of the discrepancies between HTS and microarray predictions.

We also considered the possibility that relatively low GC content of the targets could compromise hybridization-based LLMDA detection. Average log (fluorescence) intensity of probes for a given taxon strongly correlates with the average GC% of that probe set ($r=0.56$, $p = 0.0028$, $R^2 = 0.368$ for cholera victim 3090.13; $r=0.65$, $p = 2.5 \times 10^{-13}$, $R^2 = 0.653$ for plague victim 8291; **Figure 4**), but LLMDA detected taxa across the range of average log intensities. Furthermore, for taxa used for probe design, there is no significant difference in GC content between LLMDA-positive and LLMDA-negative HTS reads (two-tailed, unequal variance Student's t-test, $p = 0.252$ for 3090.13, $p = 0.779$ for 8291; **Figure 3b**). This indicates that GC content alone cannot explain a taxon's presence or absence from the LLMDA calls. Confident LLMDA log odds-based identification may also be compromised when regional preservation or amplification biases have reduced the evenness of genomic representation by the individual reads. However, for taxa with probes on the array, there is no significant difference between the proportions of unique genomic bases covered by HTS reads for LLMDA-positive and LLMDA-negative taxa (two-tailed, unequal variance Student's t-test, $p = 0.365$ for 3090.13, $p = 0.843$ for 8291; **Figure 3c**). Therefore the propensity for LLMDA detection likely derives from a complex interaction of sample composition and statistical parameters of the analysis.

Several taxa were detected only with LLMDA. This may suggest that the LLMDA is more sensitive than HTS to certain taxa, as a rarefaction analysis of the HTS data suggests that in neither sample have all the HTS-detectable families likely been observed at these sequencing depths (**Figure 5**). Cholera victim 3090.13 in particular shows a near-linear rarefaction curve, potentially explaining why it has so many more LLMDA-only calls than does plague victim 8291. However, taxa detected by both HTS and LLMDA still have significantly higher LLMDA log odds scores than taxa detected by

LLMDA alone (two-tailed, unequal variance Student's t-test, p = 0.031 for 3090.13, p = 0.013 for 8291; **Figure 3d**). This difference likely reflects the fact that LLMDA calls with smaller log-odds scores are supported by fewer detected probes, and are thus inherently less reliable. However, the relationship between log odds scores and HTS observations is imperfect, as several taxa with relatively high read counts have maximum log odds score values within the range of LLMDA-only calls (e.g., Caulobacteraceae for sample 8291 and Moraxellaceae for sample 3090.13). Again as noted above, there is no significant difference between the proportion of unique genomic bases covered by HTS reads for LLMDA-positive and LLMDA-negative taxa (**Figure 3c**). Therefore, it is likely that a more complex combination of variables drive these signal disparities.

Here we demonstrate that the LLMDA provides similar bacterial family-level metagenomic profiles of archaeological and archival specimens as HTS, especially for the most abundant families. Furthermore, as demonstrated with cholera victim 3090.13, it is potentially capable of detecting bacterial families that are insufficiently or unable to be detected even with very large HTS datasets, due to the very deep sequencing depths required to observe low abundance HTS taxa, likely common for many co-infecting pathogens. This is encouraging, since LLMDA analysis is at least one order of magnitude less expensive and labor-intensive than metagenomic HTS. As such, the technique could be productively applied in a number of research settings, depending on the specific question and the nature of the specimens. For instance, dozens of samples could be rapidly assessed for the most abundant pathogen constituents. Use of the LLMDA may also integrate well into TE-HTS studies not only by narrowing the range of targets for hybridization capture, but also by generating enriched libraries via elution from the microarray itself, which can be later sequenced. However it is clear that the profiles generated by the LLMDA and HTS are not identical, and criteria for confident family detection with both platforms remain imperfect. Though no single or simple combination of variables fully explains the signal disparities, there is good evidence that analysis techniques, GC content, and probe design drive the disagreements between the LLMDA and HTS, but that further evaluation may be able to refine these disparities. With such efforts, we expect that microarrays will evolve in the near future to become an excellent screening tool for archaeological and clinical samples where microbial profiles can be swiftly, cheaply, and accurately reconstructed, to help determine the microbial flora and its possible contribution to population health through time.

Methods

Libraries from these specimens were both shotgun HTS sequenced (divided across one HiSeq 1000 lane: 141,039,627 reads for cholera victim 3090.13, 122,830,910 reads for plague victim 8291 and utilized for LLMDA analysis. HTS datasets were compared to the NCBI RefSeq database[22] using BLAST 2.2.26+[20] and the resulting BLAST reports were parsed using MEGAN4 v.4.70.4 with the default settings.[21] Taxonomic trees were illustrated manually using FigTree (v.1.4.0; http://tree.bio.ed.ac.uk/software/figtree) based on MEGAN4 results. Indexed libraries were sent to

Lawrence Livermore National Laboratory (LLNL) for blind analysis using the 12-plex 135K Roche NimbleGen version of the LLMDA v5 array,[23] which is designed to target 3521 vertebrate-infecting species from 215 families (including bacteria, archaea, viruses, protozoa and fungi). A brief summary of the LLMDA workflow is as follows: libraries are linearly amplified via random hexamers (Cy-3 labeled) to add the necessary fluorescent signal, hybridized to the LLMDA array for 65h, washed, scanned, and analyzed. Unlike other aDNA experiments utilizing in-solution or array hybridization, "blocking oligonucleotides" were not used, as this is not a standard component of the LLMDA procedure. Arrays were analyzed using the CLiMax algorithm[13] with probe intensity threshold set to the 95th percentile of negative controls. See Supplementary Appendix for all further details.

Acknowledgements

We would like to thank David J.D. Earn, David Fisman, Joseph Tien, G. Brian Golding, Nicholas Waglechner, D. Poinar, Melanie Kuch, D. Ann Herring, Kirsten Bos and Johannes Krause, and the members of the McMaster Ancient DNA Centre for their ongoing insights and contributions to this work in the field of paleopathological research. We thank Rebecca Redfern and Jelena Bekvalac at the Museum of London Centre for Human Bioarchaeolgy for providing access to the 8291 specimen. AMD was supported by an Ontario Graduate Scholarship. HNP was supported by a CRC (Canada Research Chair) and NSERC grants. TMP was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute through the Biomonitoring 2.0 project (OGI-050).

Prepared by LLNL under Contract DE-AC52-07NA27344.

References

1 Ortner, D. J. Human skeletal paleopathology. *Int. J. Paleopathol.* **1**, 4-11, doi:10.1016/j.ijpp.2011.01.002 (2011).

2 Rizzi, E., Lari, M., Gigli, E., De Bellis, G. & Caramelli, D. Ancient DNA studies: new perspectives on old samples. *Genetics Selection Evolution* **44**, doi:10.1186/1297-9686-44-21 (2012).

3 Devault, A. *et al.* The Second Pandemic Strain of *Vibrio cholerae*, Philadelphia, 1849. *N. Engl. J. Med.* (In press).

4 Schuenemann, V. J. *et al.* Genome-wide comparison of Medieval and modern Mycobacterium leprae. *Science*, doi:10.1126/science.1238286 (2013).

5 Bos, K. I. *et al.* A draft genome of Yersinia pestis from victims of the Black Death. *Nature* **478**, 506-510, doi:10.1038/nature10549 (2011).

6 Schuenemann, V. J. *et al.* Genome-Wide Comparison of Medieval and Modern Mycobacterium leprae. *Science* **341**, 179-183, doi:10.1126/science.1238286 (2013).

7 Brogden, K. A., Guthmiller, J. M. & Taylor, C. E. Human polymicrobial infections. *Lancet* **365**, 253-255, doi:10.1016/s0140-6736(05)70155-0 (2005).

8 Gonzalez, J. M., Portillo, M. C., Belda-Ferre, P. & Mira, A. Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. *PLoS ONE* **7**, e29973, doi:10.1371/journal.pone.0029973 (2012).

9 Dabney, J. & Meyer, M. Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* **52**, 87-+, doi:10.2144/000113809 (2012).
10 Khairat, R. *et al.* First insights into the metagenome of Egyptian mummies using next-generation sequencing. *J. Appl. Genetics*, 1-17, doi:10.1007/s13353-013-0145-1 (2013).
11 Keller, A. *et al.* New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun.* **3**, doi:10.1038/ncomms1701 (2012).
12 Hoheisel, J. D. Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* **7**, 200-210 (2006).
13 Gardner, S., Jaing, C., McLoughlin, K. & Slezak, T. A microbial detection array (MDA) for viral and bacterial detection. *BMC Genomics* **11**, 668, doi:10.1186/1471-2164-11-668 (2010).
14 Erlandsson, L., Rosenstierne, M. W., McLoughlin, K., Jaing, C. & Fomsgaard, A. The Microbial Detection Array combined with random Phi29-amplification used as a diagnostic tool for virus detection in clinical samples. *PLoS ONE* **6**, e22631, doi:10.1371/journal.pone.0022631 (2011).
15 McLoughlin, K. S. Microarrays for pathogen detection and analysis. *Brief. Funct. Genom.* **10**, 342-353, doi:10.1093/bfgp/elr027 (2011).
16 Wang, D. *et al.* Microarray-based detection and genotyping of viral pathogens. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 15687-15692, doi:10.1073/pnas.242579699 (2002).
17 Palacios, G. *et al.* Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerg. Infect. Dis* **13**, 73-81 (2007).
18 Victoria, J. G. *et al.* Viral nucleic acids in live-attenuated vaccines: Detection of minority variants and an adventitious virus. *J. Virol.* **84**, 6033-6040, doi:10.1128/jvi.02690-09 (2010).
19 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, doi:10.1186/gb-2009-10-3-r25 (2009).
20 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **215**, 403-410, doi:10.1006/jmbi.1990.9999 (1990).
21 Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N. & Schuster, S. C. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* **21**, 1552-1560, doi:10.1101/gr.120618.111 (2011).
22 Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130-D135, doi:10.1093/nar/gkr1079 (2012).
23 Gardner, S., Jaing, C., McLoughlin, J. T., Be, N. & Slezak T. MDA v5 array paper. (in prep).

Table Captions

Table 1. Summary of LLMDA and HTS results

Only taxa with probes designed on the LLMDA array are shown (see Table S1 for full results). Only taxa with at least 5 reads are called with HTS-MEGAN4 analysis. Reads = number of HTS reads assigned to that taxonomic level (- = not found in HTS dataset). LO score = LLMDA log odds score (- = not called with LLMDA). Phyla abbreviations = Act, Actinobacteria; Bac, Bacteroidetes; Chla, Chlamydiae; Chlo, Chlorobi; Chl, Chloroflexi; Fib, Fibrobacteres; Fir, Firmicutes; Fus, Fusobacteria; Pro, Proteobacteria; Spi, Spirochaetes; Syn, Synergistetes; Ten, Tenericutes; The, Thermotogae; Ver, Verrucomicrobia.

Figure Captions

Figure 1. Number of bacterial families detected by HTS and/or LLMDA

Number of bacterial families (or less-specific higher taxonomic level) detected by HTS sequencing (green circles) and LLMDA analyses (blue circles). Families detected by both methods are indicated where the circles overlap. Values above the midline include all detected families, whereas values below the midline are restricted to families included in the LLMDA probe design.

Figure 2. Comparison of HTS results vs. LLMDA results. Cladograms based on NCBI Genbank taxonomy indicating results of the BLASTN/MEGAN4 HTS analysis at the family level and above compared to LLMDA results. At the leaves, circle size reflects the relative number of reads assigned to those taxa (internal node sizes only indicated if >10 reads). Colors of taxon names indicate whether that taxon had (1) reads present in the HTS data, (2) probes designed for that family on the LLMDA, and (3) LLMDA call for that taxon. Bacterial phyla and major clades are highlighted. **a. Cholera victim specimen 3090.13, b. Plague victim specimen 8291.**

Figure 3. HTS vs. LLMDA comparisons

HTS readcounts, GC content, unique genomic positions sequenced, and maximum log odds scores for both specimens plotted against whether they were detected (+) or not detected (-) with LLMDA (a-c) or HTS (d). For HTS read counts, all HTS-identified families are analyzed (a); GC content and unique genomic positions are analyzed only for families that were used for LLMDA probe design (b,c); log odds scores are only analyzed for families detected with LLMDA.

Figure 4. Average LLMDA probe GC% vs average LLMDA probe log intensity, by family

Analysis is restricted to families used for LLMDA probe design (see SOM for details). Families not detected with HTS are represented with blue triangles. Families detected wit both methods are represented with red circles.

Figure 5. HTS rarefaction analysis

Rarefaction curves showing the number of bacterial families represented by at least 5 reads as a percent of the total observed families per sample with increasing read depth (0.1% increments). Dashed lines represent lines of best fit cholera victim specimen 3090.13 is a linear curve (R^2 = 0.96936), plague victim specimen 8291 is a logarithmic curve (R^2 = 0.98217).

Table 1. Summary of LLMDA and HTS results

Only taxa with probes designed on the LLMDA array are shown (see Table S1 for full results). Only taxa with at least 5 reads are called with HTS-MEGAN4 analysis. Reads = number of HTS reads assigned to that taxonomic level (- = not found in HTS dataset). LO score = LLMDA log odds score (- = not called with LLMDA). Phyla abbreviations = Act, Actinobacteria; Bac, Bacteroidetes; Chla, Chlamydiae; Chlo, Chlorobi; Chl, Chloroflexi; Fib, Fibrobacteres; Fir, Firmicutes; Fus, Fusobacteria; Pro, Proteobacteria; Spi, Spirochaetes; Syn, Synergistetes; Ten, Tenericutes; The, Thermotogae; Ver, Verrucomicrobia.

Cholera victim specimen 3090.13				Plague victim specimen 8291			
Phylum	**Family**	**Reads**	**LO score**	**Phylum**	**Family**	**Reads**	**LO score**
Pro	Vibrionaceae	10,600	4,470.7	Pro	Enterobacteriaceae	15,062	1,640.8
Pro	Aeromonadaceae	1,877	480.0	Pro	Alcaligenaceae	11,976	880.0
Pro	Enterobacteriaceae	1,072	4,944.3	Pro	Bradyrhizobiaceae	8,189	174.4
Fir	Erysipelotrichaceae	1,039	561.7	Pro	Burkholderiaceae	7,298	10,155.0
Fir	Clostridiaceae	989	2,023.6	Fir	Clostridiaceae	5,188	1,861.8
Fir	Streptococcaceae	387	486.6	Act	Pseudonocardiaceae	4,876	474.1
Pro	Comamonadaceae	233	496.6	Pro	Comamonadaceae	3,704	466.2
Fir	Peptostreptococcaceae	216	-	Pro	Pseudomonadaceae	2,778	3,461.5
Pro	Pseudomonadaceae	178	4,313.1	Pro	Xanthomonadaceae	2,720	197.3
Pro	Moraxellaceae	122	105.2	Act	Streptomycetaceae	2,135	506.1
Pro	Xanthomonadaceae	93	228.0	Pro	Methylobacteriaceae	1,195	118.3
Pro	Burkholderiaceae	22	11,233.8	Pro	Oxalobacteraceae	1,045	119.6
Fir	Veillonellaceae	22	130.2	Pro	Neisseriaceae	903	232.0
Act	Corynebacteriaceae	19	309.5	Pro	Sphingomonadaceae	747	-
Fir	Staphylococcaceae	14	273.6	Act	Mycobacteriaceae	642	1,368.6
Pro	Pasteurellaceae	11	-	Pro	Caulobacteraceae	606	106.0
Act	Micrococcaceae	8	358.5	Pro	Acetobacteraceae	492	222.6
Pro	Neisseriaceae	8	-	Fir	Peptostreptococcaceae	324	-
Fir	Enterococcaceae	6	204.0	Act	Nocardiaceae	310	282.7
Bac	Flavobacteriaceae	6	-	Pro	Brucellaceae	274	-
Fir	Bacillaceae	5	3,077.2	Pro	Halomonadaceae	204	-
Act	Streptomycetaceae	5	523.1	Pro	Aeromonadaceae	167	-
Act	Coriobacteriaceae	5	123.6	Pro	Desulfovibrionaceae	158	218.4
Fus	Fusobacteriaceae	5	-	Fir	Lachnospiraceae	131	707.8
Fir	Paenibacillaceae	-	1,100.2	Fir	Eubacteriaceae	122	74.3
Fir	Lachnospiraceae	-	1,016.1	Act	Micrococcaceae	111	349.9
Act	Propionibacteriaceae	-	947.8	Fus	Fusobacteriaceae	99	-
Pro	Alcaligenaceae	-	745.0	Fir	Peptococcaceae	97	116.1
Fir	Lactobacillaceae	-	677.9	Act	Propionibacteriaceae	95	950.6
Pro	Desulfovibrionaceae	-	390.9	Act	Cellulomonadaceae	92	-
Act	Actinomycetaceae	-	231.2	Pro	Sutterellaceae	85	-
Act	Bifidobacteriaceae	-	225.6	Act	Gordoniaceae	84	-
Act	Micrococcineae	-	213.2	Pro	Piscirickettsiaceae	82	-
Fir	Carnobacteriaceae	-	207.6	Fir	Streptococcaceae	79	104.2
Act	Mycobacteriaceae	-	185.0	Act	Coriobacteriaceae	77	112.0
Fir	Listeriaceae	-	164.0	Act	Actinomycetaceae	74	216.8
Fir	Planococcaceae	-	157.1	Pro	Cardiobacteriaceae	70	-
Fir	Aerococcaceae	-	135.2	Fir	Lactobacillaceae	66	378.8
Pro	Deferribacteraceae	-	128.3	Fir	Veillonellaceae	65	228.7
Fir	Peptococcaceae	-	127.7	Fir	Bacillaceae	63	2,764.1
Ver	Verrucomicrobiaceae	-	127.4	Pro	Moraxellaceae	62	203.2
Act	Jonesiaceae	-	126.6	Act	Corynebacteriaceae	54	562.3
Pro	Helicobacteraceae	-	124.7	Act	Intrasporangiaceae	53	-
Pro	Caulobacteraceae	-	117.6	Act	Bifidobacteriaceae	52	748.5
Chl	Herpetosiphonaceae	-	112.8	Spi	Spirochaetaceae	52	-
Act	Brevibacteriaceae	-	112.7	Pro	Erythrobacteraceae	44	-

Act	Dermabacteraceae	-	111.3	Fir	Staphylococcaceae	43	176.7
Fir	Leuconostocaceae	-	111.0	Syn	Synergistaceae	40	108.5
Pro	Campylobacteraceae	-	107.9	Fir	Ruminococcaceae	30	100.2
Fir	Eubacteriaceae	-	95.6	Pro	Pasteurellaceae	30	80.1
Fib	Fibrobacteraceae	-	90.5	Bac	Flavobacteriaceae	25	108.9
				Pro	Vibrionaceae	24	-
				Act	Dermabacteraceae	21	108.1
				Act	Dermacoccaceae	21	-
				Act	Segniliparaceae	21	-
				Act	Tsukamurellaceae	21	-
				Spi	Leptospiraceae	20	-
				Bac	Rikenellaceae	19	-
				Fir	Erysipelotrichaceae	17	322.5
				Bac	Bacteroidaceae	17	-
				Pro	Campylobacteraceae	17	-
				Pro	Succinivibrionaceae	17	-
				The	Thermotogaceae	16	-
				Pro	Desulfomicrobiaceae	15	-
				Bac	Prevotellaceae	14	-
				Pro	Helicobacteraceae	11	120.6
				Pro	Bartonellaceae	11	-
				Ten	Mycoplasmataceae	10	-
				Fir	Leuconostocaceae	9	191.4
				Fir	Enterococcaceae	9	177.6
				Fir	Listeriaceae	9	164.6
				Bac	Porphyromonadaceae	9	-
				Pro	Legionellaceae	9	-
				Fir	Aerococcaceae	8	257.5
				Pro	Anaplasmataceae	7	-
				Pro	Coxiellaceae	7	-
				Fir	Carnobacteriaceae	6	184.4
				Pro	Bdellovibrionaceae	6	85.0
				Chla	Parachlamydiaceae	6	-
				Pro	Francisellaceae	6	-
				Pro	Rickettsiaceae	5	-
				Act	Jonesiaceae	-	108.3
				Chlo	Herpetosiphonaceae	-	101.2
				Act	Brevibacteriaceae	-	101.1
				Fib	Fibrobacteraceae	-	79.3

Figure 1. Number of bacterial families detected by HTS and/or LLMDA

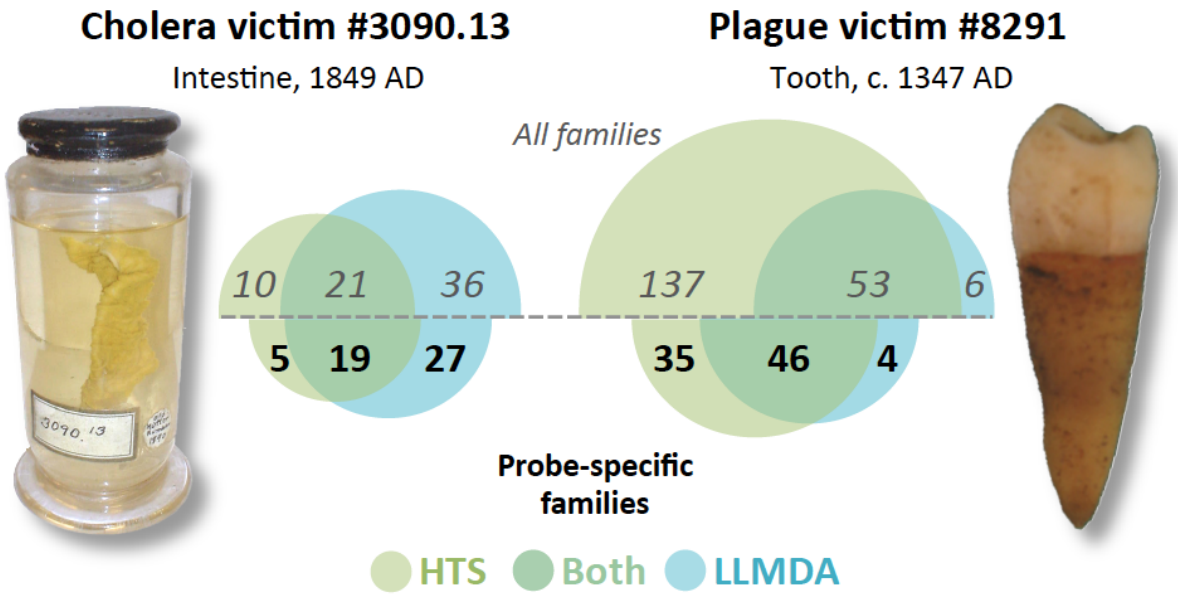
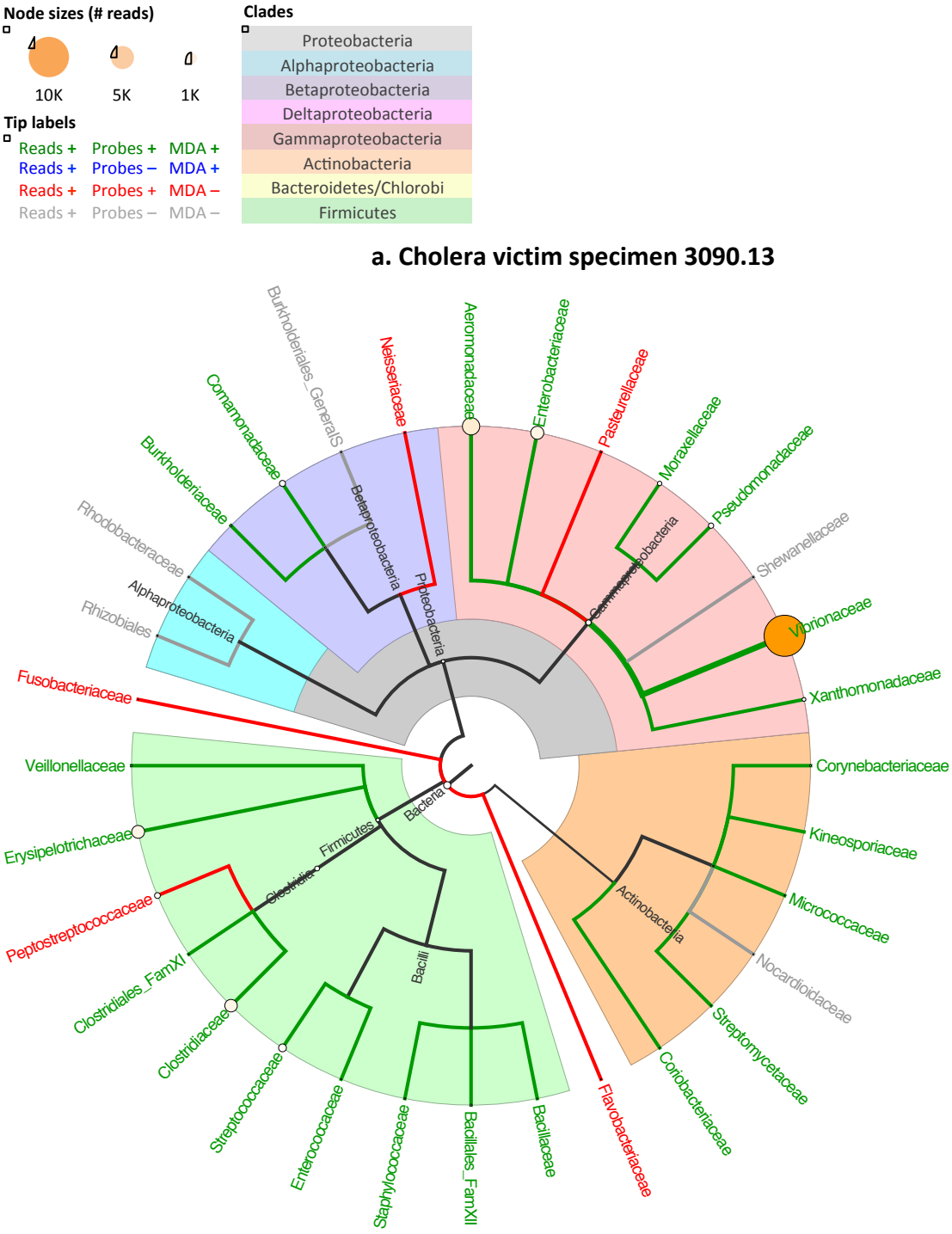


Figure 2. Comparison of HTS results vs. LLMDA results.



b. Plague victim specimen 8291

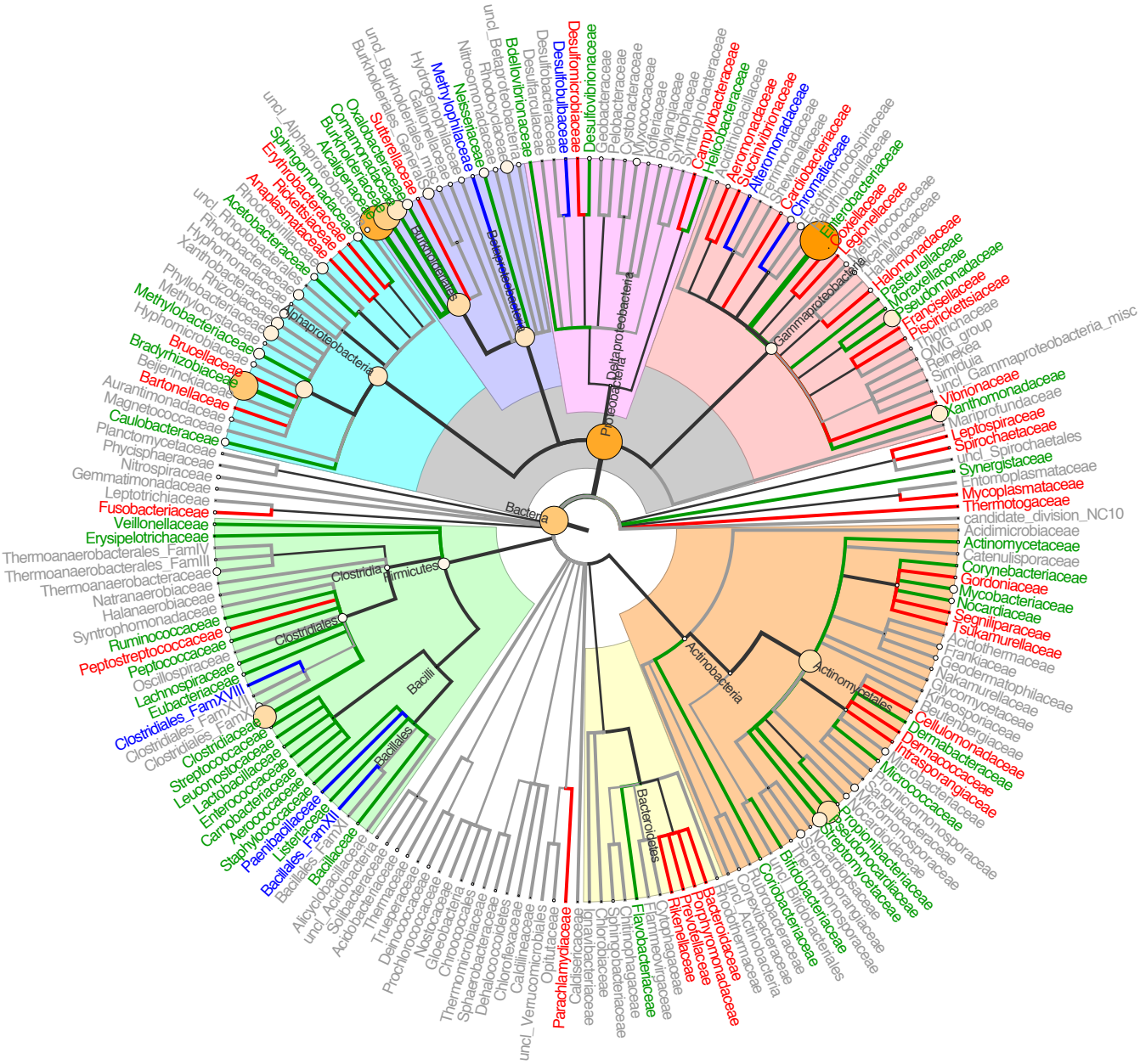


Figure 3. HTS vs. LLMDA comparisons

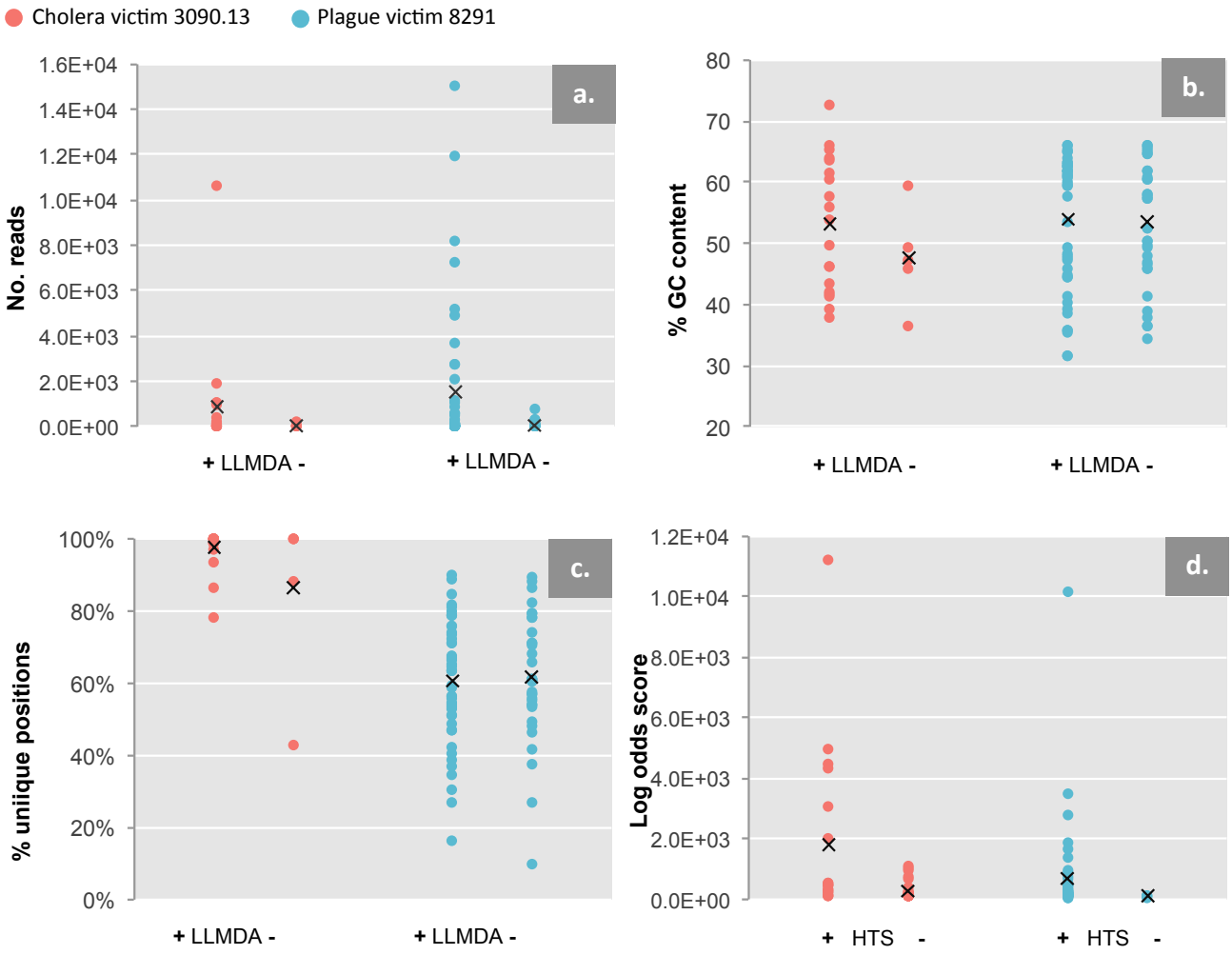


Figure 4. Average LLMDA probe GC% vs average LLMDA probe log intensity, by family

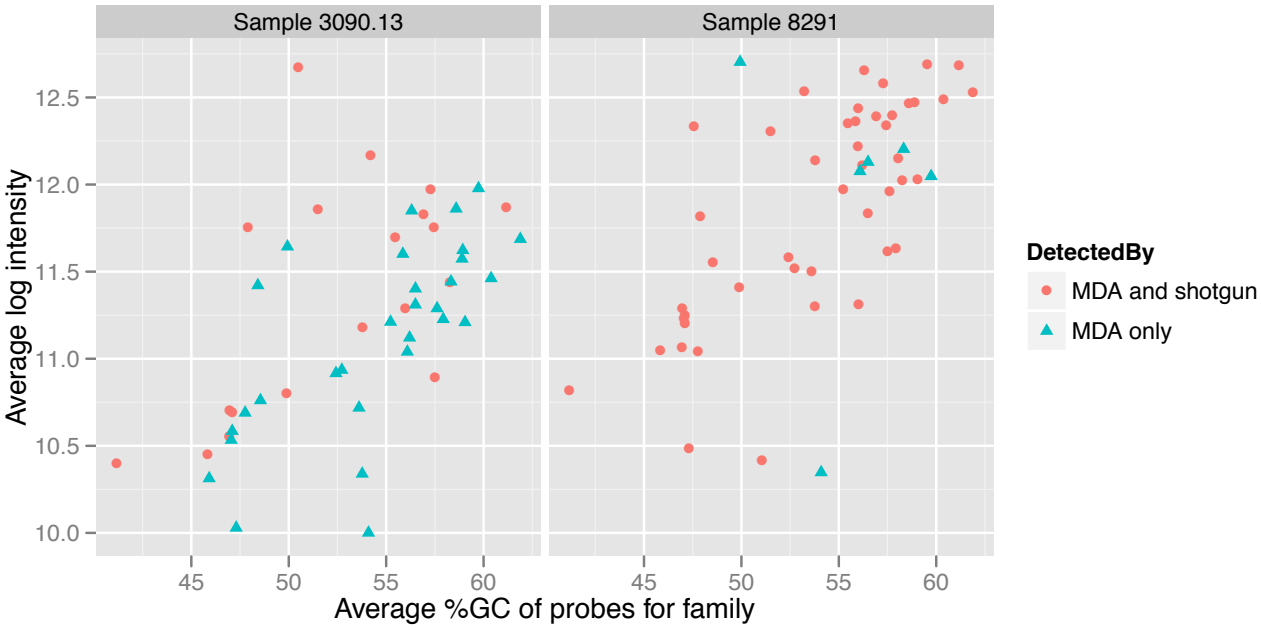
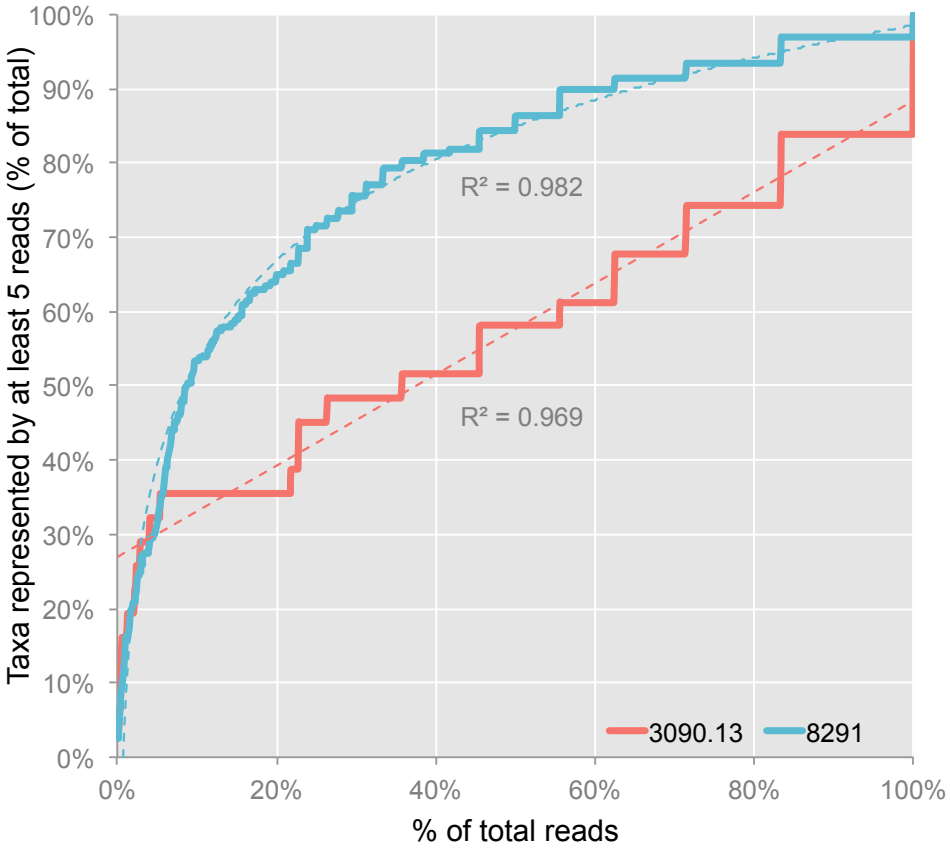


Figure 5. HTS rarefaction analysis



Supplementary Appendix for

Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array

TABLE OF CONTENTS

I. Supplemental Methods

A. Sample preparation

Sample 3090.13 is a preserved intestinal specimen from a victim of the 1849 Philadelphia cholera epidemic, sealed in a glass jar with alcohol, and stored in the collections of the Mütter Museum (Philadelphia, PA, USA). This specimen was sub-sampled, extracted, and libraries suitable for sequencing on the Illumina platform were prepared as described in reference 1. Specimen 8291 is a tooth from a victim of the Black Death buried at the East Smithfield cemetery in London in 1348-1349.[2] This specimen was sampled and extracted using the same methods as described in reference 2. Libraries suitable for sequencing on the Illumina platform were prepared just as for 3090.13 (above), as described in reference 1.

B. Shotgun HTS sequencing

Prior to sequencing, additional indexing amplification was performed in 8 reactions each sample (5 µl 0.1x diluted template DNA in 50 µl total reaction volume) of indexed library, using 400nM each indexing primer, and 11 cycles for 3090.13 and 20 cycles for 8291. The purified libraries were pooled in equimolar ratio on one lane of Illumina HiSeq 1000. Sequencing was performed by the Farncombe Family Digestive Health Research Institute (McMaster University). 100bp paired-end read chemistry was used, with one indexing read. The lane yielded 141,039,627 reads each direction from 3090.13 and 122,830,910 reads each direction from 8291.

C. Pathogen HTS assemblies

Raw R1 reads from each sample were trimmed to remove residual adaptor sequence using cutadapt (v.1)[3] with the parameters: error rate (0.16), minimum overlap (1). Reads <28bp were removed from a 24,000,000 subset of each sample, leaving 12,946,441 for 3090.13 and 12,076,222 for 8291. To calculate HTS pathogen percentages, remaining reads were aligned using *bowtie* v.0.12.7[4] with default settings to the O395 strain *V. cholerae* reference genome (NC_009456, NC_009457) for sample 3090.13 and to the CO92 strain *Y. pestis* reference genome and 3 plasmids pCD1, pPCP1, and pMT1 (NC_003143, NC_003131, NC_003132, NC_003134) for sample 8291. For 3090.13, 6,938 aligned (0.054% of reads ≥28bp, 0.029% of total reads), and for 8291, 18,931 aligned (0.157% of reads ≥28bp, 0.079% of total reads).

D. HTS BLAST & MEGAN metagenomic analysis

Raw reads from each sample were trimmed using cutadapt (v.1.1) with the parameters -b (13bp adaptor sequence), –e (errors allowed) 0, – m (minimum length, bp) 20, -q (Phred scaled quality cutoff) 20, and -O (overlap, bp) 13, leaving 118,859,751 reads from 3090.13 and 89,321,997reads from 8291 for further processing. Reads were subjected to local BLASTN-megablast analysis (v.2.2.26+) using a local copy of the refseq_genomic database (downloaded October 16, 2012), using the parameters: -task megablast, -word_size 28, -evalue 1e-10, -num_descriptions 100, -num_alignments 100. BLAST reports were parsed using MEGAN4 (v.4.70.4) using the default lowest common ancestor (LCA) parameters.[5] Full results of this analysis can be found in Table S2.

E. LLMDA Analysis

i. LLMDA v5 design

All completely sequenced genomes or elements (chromosomes, mitochondria, plasmids) as of December 20, 2011 were obtained from public sources (NCBI, J. Craig Venter Institute, etc.). These included assembled draft and finished sequences for viruses, bacteria, archaea, fungi, and the subset of protozoa known to be human pathogens or their near neighbors. These were grouped by kingdom and family. LLMDAv5 was designed using substantially the same approach as previous versions,[6] namely, finding family-specific regions in the available complete sequences, and selecting probes within those regions such that all targets are represented by both conserved and discriminating probes. The LLMDAv5 135K design has approximately 135,000 unique target probes. Conserved probes were selected favoring the most within-family conserved, thermodynamically optimal probes, so that all targets were represented by at least 15 conserved probes. Discriminating probes were selected favoring the least conserved probes for each sequence, with at least 2 per genome or sequence element. On the 135K design, only probes from families containing at least one species known to infect vertebrates were included for the viruses, bacteria, and fungi. All archaea families were included since there were few enough probes to include them all, as well as all the pathogenic protozoa previously selected for probe design. Vertebrate infecting bacterial, viral, and fungal families were selected based on literature (PubMed) and web searches to determine whether any members of a family have been found to infect vertebrates or were involved in clinical infections, and all members of a family were included even if only some of them were vertebrate-infecting. The array also included several thousand negative control probes with random sequences designed to match the length and GC% distribution of the target probes. The following numbers of species were represented: 3,521 microbial species total, including 1,856 viral species, 1,398 bacterial species, 125 archaeal species, 94 protozoan species, and 48 fungal species.

ii. LLMDA analyses

LLMDA arrays were analyzed using the CLiMax (Composite Likelihood Maximization) algorithm, described in detail previously[4], followed by some additional processing steps. We measured probe

intensities on each array using NimbleScan software (Roche NimbleGen) and reduced them to vectors of binary probe detection indicators, by comparing each target probe intensity to the 95th percentile of the negative control probe intensities. The CLiMax software processes this indicator data using a greedy iterative procedure to predict a series of targets likely to be present in the sample. In the first iteration, a target is selected by computing, for each genome in a reference target database, the log-odds of the observed probe detection data if that genome were present in the sample; the target with the highest log-odds score becomes the first element of the series. In each subsequent iteration, a conditional log-odds score is computed for each remaining target, representing the likelihood of the data if the target were added to the series, relative to the likelihood given the previously predicted targets. The target with the largest conditional log-odds score is then appended to the series. Iterations continue until there are no additional targets with positive conditional log-odds scores, meaning that no further improvement in the likelihood can be obtained by predicting additional targets.

After the initial CLiMax analysis, we filtered the list of genomes predicted to be present by rejecting those for which the array detected only a small subset of the genome regions covered by probes. In our past experience, targets with this pattern of detected probes are likely to be false positives, resulting from cross-hybridization to a similar region in another genome. Figure S3 shows examples of targets that were accepted and rejected under this filtering strategy. We aligned probes matching each selected target sequence to genome positions using BLAST. We used Gaussian kernel density estimates to approximate the positional distribution functions for all probes matching the target (with predicted detection probabilities greater than 0.85), and for the subset of these probes with intensities above the 95th percentile of negative controls, taking care to use the same bandwidth for both estimates. To quantify the difference between these two distributions, we computed the Kullback-Leibler divergence (D_{KL}) between the two density estimates. If $f_{pred}(x)$ and $f_{det}(x)$ are, respectively, the estimated density functions for the probes predicted to bind the target and the probes actually detected, evaluated at discrete positions x, then the K-L divergence is computed as $D_{KL}(f_{pred} \,||\, f_{det}) = \sum_x f_{pred}(x) \log (f_{pred}(x) / f_{det}(x))$. Targets with $D_{KL} > 4\times10^{-4}$ were removed from the predicted set; this threshold was chosen by analysis of samples of known composition, to provide a reasonable compromise between sensitivity and specificity. The numbers of target sequences predicted to be present in sample 8291 were 398 total and 204 after filtering; for sample 3090.13 the target counts were 430 total and 217 after filtering.

Finally, to enable comparison of the LLMDA results with the family-level results produced by BLAST and MEGAN analysis of HTS data, we grouped the filtered targets by family, and summed log-odds scores over targets to produce an aggregate score for each family.

II. Supplemental Results

Table S1. LLMDA and HTS analysis, full results

SEE EXCEL FILE

Table S2. HTS BLASTN/MEGAN metagenomic profiles, full results

SEE EXCEL FILE

Figure S1. Flowchart of workflow

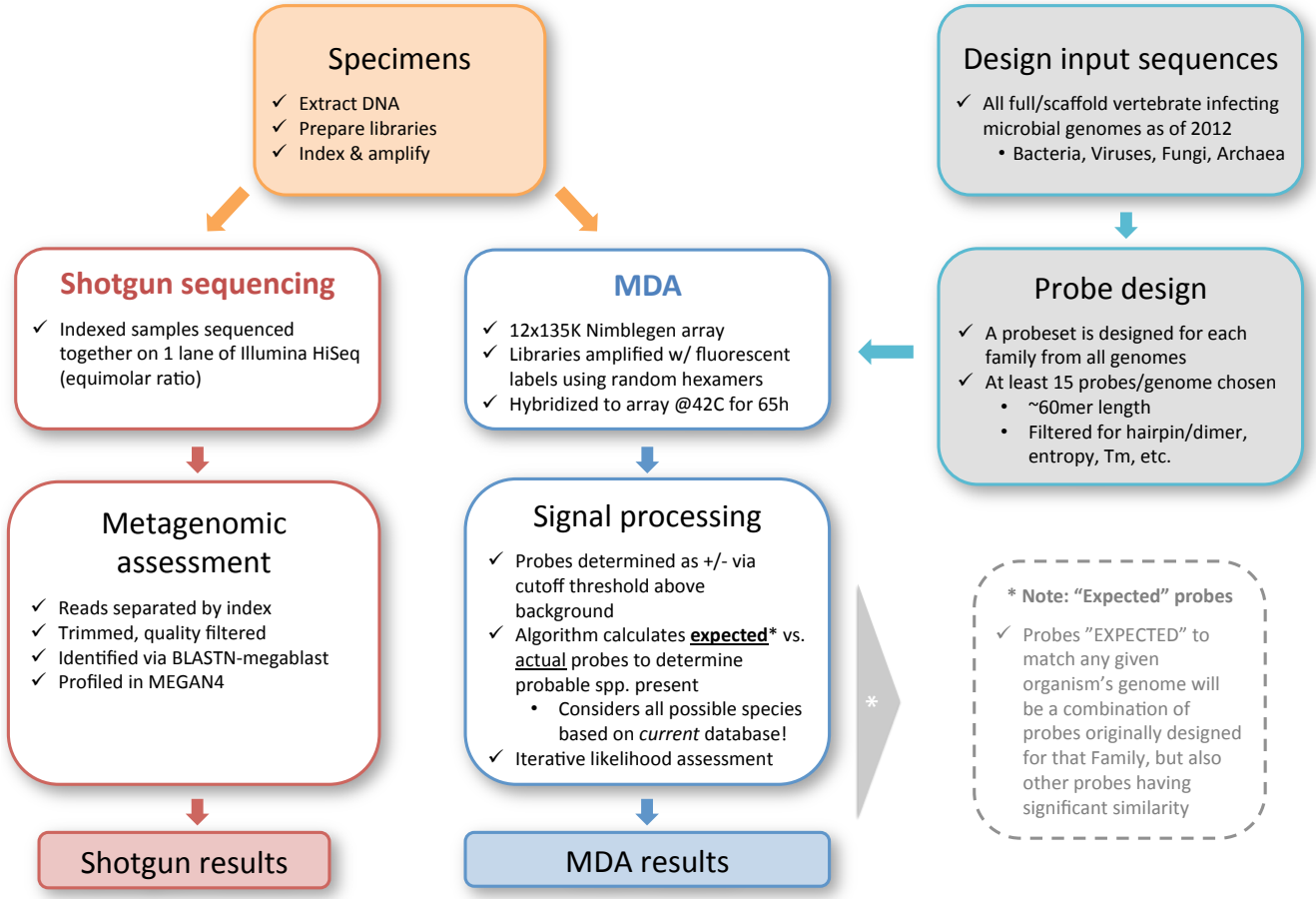
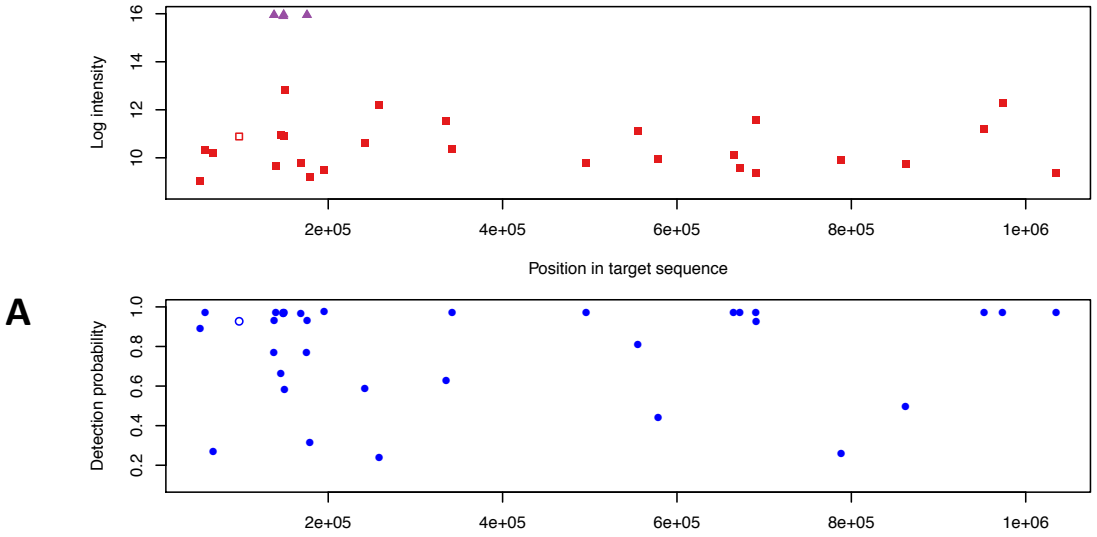


Figure S2. Number of HTS reads vs. HTS GC%, by family

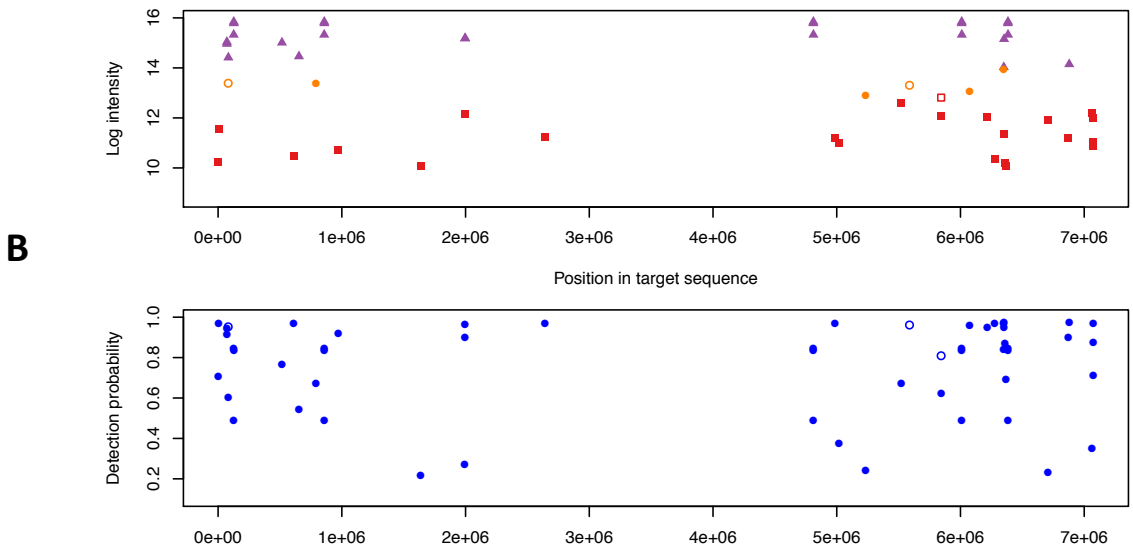
For each bacterial family detected by HTS with probes present on the MDA v5, plots of the total number of HTS reads assigned to that family versus GC% of the HTS reads (see SOM for details). Data only shown for those families with HTS representation (3090.13 = 24; 8291 = 81). Blue triangles = families not detected via LLMDA. Red circles = families detected by LLMDA.

Figure S3. Examples of LLMDA detected probe distributions

A: Log intensities vs genome position for probes targeting *Chlamydia muridarum* on array hybridized to sample 8291, and probe detection probabilities (based on similarity to target sequence) vs position. Purple triangles indicate that intensity was above the 99th percentile of the negative controls; orange circles indicated intensities between the 99th and the 95th percentiles; red squares indicate intensities below the 95th percentile. Open circles (or squares or triangles) are the probes that we excluded from the score computation, because they light up non-specifically even when there's no sample present in the hyb mixture.This target was removed from the predicted set because the only high-intensity probes came from a narrow region of the genome.



B: Log intensities and detection probabilities vs genome position for probes targeting *Pseudomonas fluorescens* Pf-5 on array hybridized to sample 3090.13. This target was included in the predicted set, since high-intensities are found from most regions of the genome that are covered by high-probability probes.



III. Supplemental References

1 Devault, A. *et al.* A 'Classical' genome for the second pandemic strain of Vibrio cholerae, Philadelphia, 1849. *N. Engl. J. Med.* (SUBMITTED).

2 Bos, K. I. *et al.* A draft genome of Yersinia pestis from victims of the Black Death. *Nature* **478**, 506-510, doi:10.1038/nature10549 (2011).

3 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17** (2011).

4 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, doi:10.1186/gb-2009-10-3-r25 (2009).

5 Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N. & Schuster, S. C. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* **21**, 1552-1560, doi:10.1101/gr.120618.111 (2011).

6 Gardner, S., Jaing, C., McLoughlin, K. & Slezak, T. A microbial detection array (MDA) for viral and bacterial detection. *BMC Genomics* **11**, 668, doi:10.1186/1471-2164-11-668 (2010).